DOCUMENT RESUME

ED 219 406 TM 820 417

AUTHOR Goodstein, H. A.
TITLE The Impact of Implementing a More Rigorous Standard for Determining the Reliability of Proficiency Tests.
PUB DATE Mar 82
NOTE 65p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 19-23, 1982).

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Basic Skills; Criterion Referenced Tests; *Educational Testing; Elementary Secondary Education; Measurement Objectives; *Minimum Competency Testing; *Standards; Test Construction; *Test Interpretation; *Test Reliability; Test Theory; Test Validity
IDENTIFIERS *Domain Specifications; Tennessee Proficiency Test

ABSTRACT

The proposed standard for judging proficiency test score reliability requires that the proportion of items passed for each objective assessed be a dependable estimate of the universe score for the domain strata established by the objective. Domain breadth is the focusing issue. Data from a field trial of the Tennessee Proficiency Test are analyzed to determine the dependability of measurements for each objective and the minimum number of items needed to assess each objective reliably. The test reports performance in each of four domains (mathematics, reading, language, and spelling) as a single summative score contrasted with the standard to describe a student as a master or non-master. The reasonableness of criterion-referenced interpretation of scores, the usefulness of empirical approaches to strengthening the power of the interpretations and implications of the findings for the construction of proficiency tests are discussed. (Author/CM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED219406

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

The Impact of Implementing a More Rigorous Standard for Determining the Reliability of Proficiency Tests

H. A. Goodstein, Ph.D.
Associate Professor of Special Education
George Peabody College
Vanderbilt University

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

H. A. Goodstein

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 820 417

A Paper Presented at the Annual Convention of the American Education Research Association, New York, March 1982.

Purpose

This paper will focus attention upon the issue of domain breadth as it relates to the development of and interpretation of scores from minimal competency or proficiency tests. Specifically, I will examine a basic skills proficiency test that includes multiple objectives with a limited number of items (four) to assess each objective. This test reports performance in each of four curriculum/content domains (mathematics, reading, language, and spelling) as a single summative score to be contrasted with a standard in order to describe the student as a master or non-master. At issue will be the question as to the reasonableness of "criterion-referenced" interpretation of such scores and the usefulness of empirical approaches to strengthening the power of those interpretations.

Definition of a Criterion-Referenced Test

Hambleton (1980) has suggested that the following definition of a criterion-referenced test is probably the most widely used:

> A criterion-referenced test is constructed to assess the performance levels of examinees in relation to a set of well-defined objectives (or competencies) (p. 421).

In elaborating upon this definition, Hambleton observed that terms such as objectives, competencies and skills may be used interchangeably. The content or behaviors defining each objective should be clearly described. Also, when more than one objective is measured in a test, usually the test items are organized into non-overlapping subtests corresponding to the objectives, and examinee performance is reported on each of the objectives.

Domain Breadth

While criterion-referenced measurement has emerged as a multi-faceted concept (Nitko, 1981), its original purpose was to more effectively evaluate learning as a result of direct instruction (Glaser, 1963). Baker (1981) has recently reviewed the origins of the measurement technology in instructional psychology and curriculum development. The relationship between instruction and measurement was simple and direct. Instructional objectives, often expressed as "behavioral" objectives (Mager, 1962), fixed the domains for both instruction and measurement. Early efforts in the development of criterion-referenced tests focused upon very narrowly defined domains reflecting contemporary practices in instructional psychology.

The introduction of the term domain into the lexicon of criterion-referenced measurement is often attributed to the efforts of Hively and his associates (Hively, Maxwell, Rabehl, Sension, and Lundin, 1973) to provide "complete" specifications of a measurement domain. It should be noted that these efforts were directed toward further stratification of rather precise instructional objectives. Examination of the suggested item forms reveals a commitment to single-objective domains (the specificity of the objective consistent with contemporary instructional psychology) with the sampling plan explicated for variations within the domain fixed by the objective.

More recently, instructional planning practices have encouraged the use of more broadly stated objectives (Kapfer, 1978). These practices evolved as it became clear that the exponential expansion of Mager-type objectives could not be maintained nor managed in curriculum planning beyond the early skill acquisition stages of the elementary school curriculum. Popham & Lindheim (1980) describe one such approach to the broadening of

objectives as broad scope objectives (instructional targets) with each objective subsuming several enroute, enabling skills, all of which potentially could be measured (Mappus, 1981).

As the range of skills subsumed by a single objective has broadened, so has the concept of a measurement domain. While we continue to use the term domain, it appears that its referent is no longer the construct operationalized by Hively, et al (1973). The question of breadth or complexity of a domain has emerged as a significant validity issue for criterion-referenced measurement (Linn, 1980). One reason may relate to the weakening of a professional concensus over the level of specificity defining a well-formulated instructional objective.

The framework of behavioral analysis shared by most measurement professionals suggest that any universe of behavior or content can be analyzed into succeedingly smaller levels of subtasks. Conversely, any set of skills at one level of specificity can be subsumed or aggregated into a broader skill at the next higher level of specificity. At this time, there exists little guidance as to the level of specificity of an instructional objective (or broad scope instructional target) required to provide the "well-defined" domain essential for a criterion-referenced interpretation of the test scores.

Hambleton (1981) has pointed out that the essential feature of a criterion-referenced test is clarity of definition rather than breadth or complexity. If we accept Hambleton's definition of a criterion-referenced test as requiring non-overlapping measurement of objectives and a lack of standards for the level of specificity of objectives, what we require is an operational definition of "clarity of definition".

Most current methodologies seek to operationalize clarity of definition by providing layers of stratification to instructional objectives often written at broader (less precise) levels of specificity. This is an effort to approach the idealized model of domain-referenced measurement, where detailed, explicit rules could generate a finite pool of items (Berk, 1980b). However, in the absence of a-priori standards for determining the depth of the analysis process or the range of variables to be considered, it is difficult to evaluate or judge the products (the measurement domains) as to clarity of definition. Obviously, broader domains require consideration of more variables and invite more controversy over the selection of variables referred to in the analysis (Quelmallz, 1981).

Linn (1980) also has suggested that the breadth of coverage of a criterion-referenced test need not be narrow. However, Linn has observed that most current models or methodologies for providing clarity of domain definition that result in increased specificity also result in very narrowly defined domains (Linn, 1980).

Popham (1978) has argued that empirical tests of item homogeneity (functional homogeneity) were overly restrictive in determining membership of a collection of items within a domain and, thus, were not useful in judging clarity of domain definition. Rather, he supports a less restrictive test of derivative homogeneity; that is, that the items are included in the domain if they are judged as derivative from a set of item specifications. This position is consistent with the movement toward broader objectives, broader domains, and the likelihood of variance in item difficulty as more variables (identified and unidentified) influence performance. Baker (1981) has observed, "as strict expectations for item homogeneity faded, so have guidelines for the restrictiveness of domains" (p. 9).

While the item specifications proposed by Popham provide a heuristic process for clarifying the definition of a selected domain, they offer no solution to the issue of breadth or complexity of the selected domain.

Domains of Analysis for Proficiency Tests

In recent years, an alternate application of criterion-referenced measurement technology has been in the development of proficiency tests or tests of minimal competency. Such tests have increasingly been used to determine eligibility for the receipt of a high school diploma. Many of these tests include multiple objectives organized into sections or subtests defined by broad curriculum domains (e.g., mathematics or reading). Typically, a proportion correct score is calculated from performance on all items (across all objectives) in the curriculum domain. This score is then contrasted with a standard or cut-off score in order to make an evaluative classification (Fitzpatrick, 1981) in regard to competence or proficiency. In some instances, performance by objective results are also reported to provide "diagnostic" or remedial guidance for students not attaining the standard.

The "domain of analysis" (similar to unit of analysis in experimental design) for most proficiency tests are the content or curricular domains defined by a set of instructional objectives. Note I am now using the term domain in this context in a broader sense than suggested by the work of Hively, et al. (1973). These objectives are selected on the basis of being "important", "relevant", and/or "representative" often as the result of an extensive review process involving both professional educators and the general public. Often detailed test item specifications are developed to operationalize each objective in order to develop or select a derivatively homogeneous set of test items. Thus, the domain is defined by a collection

of objectives, each represented as well-defined in regard to the stratum or subdomain that the objective represents.

Psychometric analyses including, but not limited to, reliability studies of this genre of proficiency tests are typically restricted to the domains of analysis, the content or broad curriculum domains. Restricting the analyses to these aggregate domains represents a tacit acceptance of the premise that those domains are sufficiently well-defined to allow for a strong criterion-referenced interpretation of test scores and decisions made as a function of those scores.

Construct Validity

Interpretation of all test scores requires consideration of construct validity (Messick, 1980; 1981). Construct validity issues have been given more attention recently by measurement professionals concerned with appropriate interpretation of criterion-referenced test scores (Haertel, 1981; Linn, 1979b; 1980; 1981; Messick, 1980; 1981; Shoemaker & Johnson, 1981) and decisions made as a result of those scores (Fitzpatrick, 1980; 1981; Hambleton, 1980b). Linn (1980) has noted that any interpretation of a score that goes beyond a simple empirical summary involves inference. Inferences regarding the meaning of test scores always involve constructs and need to be supported by evidence and logical analysis (Haertel, 1981; Linn, 1980).

A typical minimal level of inference traditionally associated with criterion-referenced tests is use of the proportion correct score as an estimate of a domain score. Most frequently the domain score is conceptualized as a true score (classical test theory) or a universe score (generalizability theory) based upon the concept of an item-sampling model (Linn, 1981; Wilcox, 1980). Setting aside important construct validity issues in regard to the distinction between predicted levels of item performance and capability for performance (Linn, 1981), distinctions between items and responses (Fitzpatrick, 1980) and

distinctions between items and skills (Haertel, 1981; Wilcox, 1980); one is still left with the need to assess the adequacy of the popular item-sampling model to allow minimal level inferences regarding performance on a multiple objective proficiency test.

Linn (1980) has noted that the major hurdle in using an item-sampling model is in defining the universe in such a way that sampling requirements can be satisfied. To the extent that specification of the item domain could approach the complete listing of all items, a random sample of such items should yield scores that are unbiased (maximum likelihood) estimates of the domain universe scores (Linn, 1981). To the extent that the ability to completely explicate the domain and/or assure random sampling of items from the domain is compromised, estimates may be biased and interpretative power diminished.

The problem is explication of the item domain and representative sampling of items from the domain (while troublesome at the objective level) compounds at the content or curricular domain level. The domain universe could be conceptualized as stratified by the measurement objectives and "weighted" similar to the distribution of items per objective on the proficiency test (Cronbach, 1980). Thus, if the proficiency test equally weighted each objective by measuring it with the same number of items, it follows that if one were to randomly sample items from the domain test item universe, one would be equally likely to select a test item for any objective within the domain.

If one rejects this conceptualization of a hypothetical domain test item universe as unreasonable, how does one interpret scores on such a proficiency test? One could argue that the domain is sufficiently ambiguous (unclearly defined) that the test score has no legitimate criterion-referenced interpretation. Rather, the score represents a level of performance on a broader construct (e.g., reading or mathematics achievement) and the test should be evaluated as to its reliability and validity as a measure of that broader

construct. Some would argue that standards could be established and decisions made as to the level of the construct. Others might be more hesitant, now that reference to specific "skills" or "competencies" is replaced in public rhetoric with more common terms such as reading achievement.

I support a third conceptualization that implies that the number of items established to assess each objective reliably estimates the domain facet score for the stratum corresponding to each objective. Thus, regardless of the number of items in each stratum of the domain universe, the number of items to assess each objective is sufficient to estimate with minimal error performance in the facet of the universe defined by the stratum. The total score is the unweighted average of performance across strata. A typical performer might be expected to perform on any of the objectives at a performance level best estimated by the performer's total score. If average difficulty of various objectives varies (as it does in practice), the expectation would modify to the relative position of a typical performer on the measurement scale for each objective.

The implication of such an interpretation is that a student's performance on the set of items measuring each objective would not be unduly impacted by errors of measurement. Alternatively stated, the assumption inherent in this interpretation is that each objective included in the test is assessed at an acceptable level of reliability by the four items measuring that objective.

What is being argued is that the typical "domain of analysis" used with the interpretation of criterion-referenced proficiency tests will not be sufficiently well-defined in the absence of additional evidence that each of the objectives that fixes the stratum from which items are drawn can

be measured within tolerable levels of measurement error. The impact of this more rigorous criteria for defining the reliability of a typical proficiency test is the focus of the data analysis to be reported.

Definition of Reliability

The term "reliability" conveys various meanings and implications in regard to norm-referenced measurement and thus has contributed to a confusing literature in the area of criterion-referenced measurement (Traub and Rowley, 1980). Therefore, it is necessary to define the use of the term in the context of this study. Applying the classification system suggested by Traub and Rowley (1980), the measurement assumptions are that the types of skills being measured are continuous (as opposed to binary) and that the intended use of the test is for decision-making (as opposed to measurement). As opposed to alternative concepts of loss function (threshold or linear), squared-error loss function will be used to estimate seriousness of classification errors.

The approach to the estimation of reliability in this study is an application of Brennan's extension of generalizability theory to domain-referenced testing (Brennan, 1978; 1980) using GAPID (Brennan, 1979), a computer program for performing generalizability analyses for single factor designs.

Linn (1979a) has noted the affinity between item sampling models used in the interpretation of criterion-referenced test scores and generalizability theory, and has suggested its appropriateness in the evaluation of competency tests.

Brennan (1978; 1980) proposed a dependability coefficient (Phi) that used an alternative definition of measurement error derivative from generalizability theory appropriate for tests that are perceived to be item samples

from a well-defined domain. One may compute Phi coefficients for any predetermined standard or cutscore (lambda). This follows the logic first proposed by Livingstone (1972) substituting the concept of measurement error derived from generalizability theory for that associated with classical test theory (Traub & Rowley, 1980).

The Phi (lambda) coefficients estimate the dependability of the test at a predetermined cutscore. These coefficients increase as the distance between the cutscore and the observed proportion correct mean score increases. The lowest level of Phi over a range of lambdas is where lambda would be equal to the population mean. Brennan and Kane (1977) have referred to this coefficient as a general purpose index of dependability and suggest it always be computed for any series of reliability studies.

Brennan's Phi coefficient at its lowest level can be shown to be mathematically equivalent to KR-21 for dichotomously scored items (Brennan, 1977), the Phi coefficient being derived from generalizability theory while KR-21 being derived from classical test theory (Traub & Rowley, 1980).

The selection of the approach for estimation of reliability was as much a pragmatic decision (availability of an efficient computor program) in contrast to a purely methodological decision. While the investigator believes generalizability theory provides a strong psychometric basis for criterion-referenced measurement, alternative psychometric models could have been selected without jeopardizing the goals of the data analysis.

The investigator was cognizant of the psychometric problems inherent in estimating reliability from a limited set of test items. Hambleton, Swaminathan, Gifford, & Mills (1981) recently estimated reliability of two-item tests using the method developed by Subkoviak (1976) which is based

upon threshold loss function. While acknowledging that the estimates of reliability were high to an unknown extent because of a short-coming of the Subkoviak method with short tests, the authors suggest that two-item tests may be expected to reliably assess narrowly defined objectives. Thus, it would appear that no a-priori limits exist in practice to define minimum number of items for measurement of an objective for a criterion-referenced test.

Results

The data to be reported on are the item responses of samples of Tennessee public school students in the sixth, seventh, and eighth grades to one or more of the subtests of the 1981 Tennessee Proficiency Test - Form IJ. The 1981 Tennessee Proficiency Test - Form IJ consisted of four sections or sub-tests -- mathematics, spelling, language, reading. The total test samples 50 of the 80 basic skill objectives listed as required proficiencies in the State Board Rules, Regulations, and Minimum Standards (Tennessee State Department of Education, 1981). The 50 objectives are distributed within the test by section as follows: mathematics - 19; spelling - 6; language - 14; reading - 11. A listing of short descriptive phrases for the Form IJ objectives is contained in Appendix A. Each objective is assessed with four test items resulting in a total test of 200 items distributed as follows: mathematics - 76 items; spelling - 24 items; language - 56 items; reading - 44 items.

The data were acquired in an effort to assist a panel of consultants in recommending standards to the State Department of Education (and ultimately to the State Board of Education) to be used in certifying eligibility for high school graduation. Also, the data were used to guide modification of test

item specifications and test item reviews for the next year's version of the test.

School districts were identified where data from norm-referenced testing programs indicated that average performance in the relevant content domain was at grade level for one or more of the three grade levels selected. Students in the school district were administered the subtest(s) of the 1981 Tennessee Proficiency Test - Form IJ indicated by the criteria. Note that this test administration plan resulted in different compositions of students taking each subtest across each of the three grade levels. Also, while the school district on average was at grade level, the students administered the proficiency tests ranged the entire gamut of achievement.

The data were transmitted to the investigator in the format of a person by item matrix stratified by grade level and content domain. All items were dichotomously scored and coded. Omitted item responses (a quite limited number) were scored as incorrect and no correction for guessing adjustments were considered in the data analyses.

Table 1 indicates the number of students tested at each grade level for each section of the test. Sample sizes ranged from 223 students for the

---

Insert Table 1 about here

---

language section at grade 7 to 719 students for spelling at grade 6. Across grade levels, total sample sizes ranged from 977 students for language to 1368 students for spelling.

Table 2 indicates the average difficulty of each section of the test for each grade level. The mathematics subtest was considerably more diffi-

cult than the other subtests, with P values ranging from .58 for grade 6

Insert Table 2 about here

to .72 for grade 8. The language and spelling subtests had P values exceeding .80 at each grade level and demonstrated minimal lessening of difficulty at succeeding grade levels.

Table 3 indicates the dependability estimate (Phi coefficient) for each section of the test. Dependability estimates for the mathematics, language,

Insert Table 3 about here

and reading subtests all exceed .85, with the coefficients for the mathematics subtest exceeding .90. Dependability estimates for the spelling subtests are more marginal, ranging from .81 to .85.

As indicated previously, the "domain of analysis" selected for this study was the individual objective with the associated four test items. Accordingly, dependability estimates were made for each of the 50 objectives at each grade level. Both Phi coefficients and Phi (.75) coefficients were calculated.

The rationale for generating Phi (.75) coefficients was that for the single objective domains, there was an increased likelihood for restricted variance in the sample due to the lack of difficulty of a particular objective. Such objectives have reduced Phi coefficients when lambda is equal to the mean. In practice, however, such objectives would be dependable in describing masters and nonmasters with a typical standard set at a lower level and substantially below the observed proportion correct. For this study, lambda was set at .75. This standard was selected arbitrarily (Glass, 1978)

on the basis of its proximity to typical standards for proficiency tests and its ease of interpretability for four item tests (3 out of 4 correct).

Appendix B contains a tabular presentation of P values, Phi coefficents and Phi (.75) coefficients for each objective at each grade level. Examination of this data revealed relative consistency of the dependability estimates for an objective across grade levels. As a result, it was determined to collapse the data across grade levels and perform the remaining analyses on the aggregate samples.

Appendix C contains a tabular presentation of P value (average P value for the four items assessing each objective), range of item P values, range of item-objective biserial correlations, Phi coefficient, and Phi (.75) coefficient for each objective. Additionally, the number of items is indicated that would be required to increase Phi (.75) to equal or exceed .85. A level of .85 was selected as a minimal level of dependability to be achieved for use in individual decision-making.

Prophecy D studies were conducted to estimate the number of items required to achieve that level of dependability. The remaining discussion of results will refer to the data in Appendix C.

The dependability analysis of the four-item objective tests revealed substantial differences among the four subtests. This analysis is displayed in Table 4. The mathematics section was composed of 10 objectives for which dependability coefficients exceeded .85, five of which exceeded .90. Only

---

Insert Table 4 about here

---

two of the 19 objectives had dependability coefficients of less than .70 for

the four item tests. In contrast, only one spelling objective, one reading objective, and no language objective could be dependably assessed at a level exceeding .85 by the four item tests. The majority of language objectives (9 of 14), reading objectives (9 of 11) and one-half of the spelling objectives (3 of 6) could not be assessed at a dependability level exceeding .70 by the four item tests.

Table 5 summarizes the results of the prophecy D studies for the objective level tests for each of the sections of the 1981 Tennessee Proficiency Test - Form IJ. For the mathematics section, applying the more rigorous standard for test length in order to reliably assess each objective would

---

Insert Table 5 about here

---

result in an additional 36 items, increasing the length of the section from 76 to 112 items or an increase of 47%.

For the spelling section, applying the more rigorous standard for test length in order to reliably assess each objective would result in at least an additional 39 items (prophecy D studies were terminated if 20 items did not raise Phi (.75) to the designated .85 level), increasing the length of the section from 24 to 63 items or an increase of 163%.

For the language section, applying the more rigorous standard for test length in order to reliably assess each objective would result in at least an additional 143 items, increasing the length of the section from 56 to 199 items or an increase of 392%.

For the reading section, applying the more rigorous standard for test length in order to reliably assess each objective would result in at least

an additional 131 items, increasing the length of the section from 44 to 175 items or an increase of 300%.

Discussion

The results of the forecasting studies suggest that with the exception of the mathematics section of the proficiency test, the impact of the more rigorous criterion (with the current collection of objectives) would result in increasing the number of items in each section beyond reasonable limits of practicality.

Items for the Tennessee Proficiency Test were developed from item specifications similar to those suggested by Popham (1978). While most of the test items would be judged congruent with the instructional objective and derivatively homogeneous with the item specifications, sets of test items varied considerably in regard to variance in item difficulties and biserial correlations. Obviously application of a consistent structure or process for the explication and definition of item domains does not result in equivalent levels of homogeneity across different content or curricular domains.

Diversity in the difficulty levels of items within a domain can be expected to vary according to the number of variables related to performance within the domain and their interactions with individual items. If we conceive of our tests as samples of items that could potentially populate the domain, domains composed of homogeneous items (in a functional sense) should require fewer items to achieve reliability or dependability in estimating domain performance than domains composed of less homogeneous items.

In one sense, the reliability estimates for any objective might be used as an empirical guide for determining the number of items to be developed in order to make a reliable inference regarding domain (stratum) performance.

Linn (1979b) has pointed out the close correspondence between the notions of generalizability theory and those of content validity. Linn preferred to discuss generalizability concepts under the rubric of content validity. Since "as commonly thought of, reliability need not involve the careful attention to unambiguous definition and sample representativeness" (Linn, 1979b, p. 117). The intent herein is to involve such attention into our reliability studies.

Berk (1980a) has observed that the problem of determining appropriate test length can be approached from a practical perspective based upon research evidence (e.g., Berk, 1980c) or from a purely technical perspective (e.g., Wilcox, 1980). Our view is consistent with the perspective of Berk (1980c), with the requirement that the nature of the inferences to be made from test scores be considered prominently as a factor in determining test length. If the nature of those inferences explicitly or implicitly consider reference performance at the objective level, consideration of test length should first focus at that level, prior to consideration of the length of the total test. For example, the current description of the Basic Skills Testing Program in Tennessee (Tennessee State Department of Education, 1981) indicates that "Four items or questions were used to test a student's understanding or mastery of each objective," (p.3).

Empirical considerations in isolation cannot determine number of test items to be included on a criterion-referenced proficiency test. The total content domain scores for the Tennessee Proficiency Test were reasonably reliable. It is a common observation that sufficient test length will produce high reliability estimates for a collection of technically acceptable test items reflective of a common factor regardless of the diversity of actual

items in a domain. This is observed in most carefully constructed norm-referenced achievement tests. However, I have argued that if the intended inference is in regard to skills or instructional objectives mastery, the demonstration of stability in the total test score is insufficient data to support the inference we desire to make regarding test performance.

Shoemaker & Johnson (1981) concluded from their study of construct validity of a multiple-objective criterion-referenced mathematics test, "many of the expectations postulated in the literature for criterion-referenced tests may be valid at the objectives level, but not at the total test level." (p. 12) Their nomological network suggested that the total test scores might be most appropriately interpreted as an estimate of general mathematics achievement (similar to comparable norm-referenced test scores). This raises serious questions as to whether broad, multiple-objective domains, are sufficiently well-defined to provide valid criterion-referenced interpretations.

Popham & Lindheim (1980) have observed "that a skill that is too broad in its scope will never prove serviceable as a focus for criterion-referenced measurement." (p.4). Reading comprehension was given as an example of a skill that is too broad in scope. However, examination of the broad-scope objectives developed by IOX for the South Carolina Department of Education reveals the following broad-scope objectives for Mathematics:

1. CONCEPTS: The student can apply numerical concepts.
2. OPERATIONS: The student can compute accurately.
3. MEASUREMENT: The student can apply measurement concepts.
4. GEOMETRY: The student can apply geometric concepts
5. PROBLEM/SOLVING: The student can solve problems requiring the the use of mathematics (Mappus, 1981, p.5).

While these objectives might be viewed as somewhat more narrow than "the student will comprehend written text," the differences may exist only in the eye of the beholder.

Messick (1981) has recently commented on the implications of the compound nature of the interpretations we make of test scores in minimal competency testing programs:

> Since the overlay construct theories of mastery and minimum competency are at least one step removed from the primary trait or skill theories underlying test development, it is not always readily apparent which features of the overlay construct should be addressed in the construct validation of the compound interpretation. This kind of nomological looseness tends not only to undercut meaningful test interpretation, but, by the same token, also tends to subvert responsible test use. (p.17)

If one were to construct proficiency tests with the more rigorous criterion of strata (objectives level) reliability influencing test construction practices, several possible outcomes are suggested. Under one scenario, the number of items developed to measure each objective would vary in relation to the presumed diversity of test items within the stratum. Some objectives might be deleted or modified in response to low reliability estimates that persisted in spite of an increase in the number of test items used to measure the objective. The resulting tests would probably focus upon fewer objectives. As more items were added to measure the broader objectives, practical considerations of desired total test length would provide an impetus to the elimination of "marginal" objectives.

Under a second scenerio, objectives would be consolidated or refined in an attempt to achieve a degree of homogeneity might be judged by a consistent

range of reliability estimates among all objectives (each being measured by the same number of items). This would probably result in an increase in the number of test items for each objective over the limited number currently used on some proficiency tests. It could be accomplished through the consolidation of narrower stratum or the refinement of more diverse stratum.

Under either scenario, the net impact of the more rigorous criterion would be to increase the number of items for many of the objectives and a probable reduction in the total number of objectives assessed.

Performance on these modified tests could be reported reliably at two distinct levels. Total subtest performance within a domain could be compared with a standard to make a classification decision. For those whose performance didnot achieve this standard, reliable performance data could be reported for each objective suggesting necessary remediation.

The modified tests could alternatively be scored on an objective attainment basis, either originally or upon retests adapted to the performance deficits of the students failing to achieve the original total score standard.

What measurement specialists desire are proficiency tests that are structurally different from typical norm-referenced achievement tests. The process of selecting instructional objectives, writing test items congruent with detailed item specifications, and imposing a standard on a test is insufficient evidence that the scores of the product can have a legitimate criterion-referenced interpretation.

Earlier criticism of the South Carolina broad-scope objectives was offered in the context of defining limits on domain breadth. However, the process of test development pursued by the South Carolina State Department of Education (Mappus, 1981) must be commended in its attempt to evolve a

series of tests that are indeed structurally different than norm-referenced achievement tests. Hopefully, we will have the opportunity to study alternative processes for test development and conduct the reliability and validity studies necessary to support responsible test use and interpretation.

REFERENCES

Baker, E. L. A multi-disciplinary review of criterion-referenced measurement. Paper presented at the annual meeting of the American Educational Research Association, Boston, April 1980.

Berk, R. A. A framework for methodological advances in criterion-referenced testing. Applied Psychological Measurement, 1980, 4, 563-573. (a)

Berk, R. A. Domain-referenced versus mastery conceptualization of criterion-referenced measurement: A clarification. Paper presented at the annual meeting of the American Educational Research Association, Boston, April 1980. (b)

Berk, R. A. Practical guidelines for determining the length of objectives-based, criterion-referenced tests. Educational Technology, 1980, 20, 36-41. (c)

Brennan, R. L. KR-21 and lower limits of an index of dependability for mastery tests (ACT Technical Bulletin No. 27). Iowa City, IA: American College Testing Program, 1977.

Brennan, R. L. Extensions of generalizability theory to domain-referenced testing (ACT Technical Bulletin No. 30). Iowa City, IA: American College Testing Program, 1978.

Brennan, R. L. Handbook for GAPID: A fortran IV computer program for generalizability analyses with single-facet designs. (ACT Technical Bulletin, No. 34). Iowa City, IA: American College Testing Program, 1979.

Brennan, R. L. Applications of generalizability theory. In R. A. Berk (Ed.), Criterion-referenced measurement: The state of the art. Baltimore, MD: The Johns Hopkins University Press, 1980.

Brennan, R. L. & Kane, M. T. An index of dependability for mastery tests. Journal of Educational Measurement, 1977, 14, 277-289.

Cronbach, L. J. Validity on parole: How can we go straight? New directions for testing and measurement. San Francisco, CA: Jossey-Bass, 1980, 99-105.

Fitzpatrick, A. R. The meaning of content validity. Paper presented at the annual meeting of the American Educational Research Association, Boston, April 1980.

Fitzpatrick, A. R. Validating decisions made with criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, April 1981.

Glaser, R. Instructional technology and the measurement of learning outcomes: Some questions. American Psychologist, 1963, 18, 519-521.

Glass, G. V. Standards and criteria. Journal of Educational Measurement, 1978, 15, 237-261.

Haertel, E. A theoretical framework for defining what is measured by criterion-referenced tests. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, April 1981.

Hambleton, R. K. Contributions to criterion-referenced testing technology: An introduction. Applied Psychological Measurement, 1980, 4, 425-446. (a)

Hambleton, R. K. Test score validity and standard setting methods. In R. A. Berk (Ed.) Criterion-referenced measurement: The state of the art. Baltimore, MD: Johns Hopkins University Press, 1980. (b)

Hambleton, R. K. Advances in criterion-referenced testing technology. In C. R. Reynolds & T. B. Gutkin (Eds.) Handbook of School Psychology. New York, NY: Wiley, 1981.

Hambleton, R. K. Swaminathan, H., Gifford, J., & Mills, C. Individualized criterion-referenced testing technical manual. Tulsa, OK: Educational Progress Corporation, 1981.

Hively, W., Maxwell, G., Rabehl, G., Sension, D., & Lundin, S. Domain-referenced curriculum evaluation (CSE Monograph Series in Evaluation, No. 6). Los Angeles, CA: University of California, Center for the Study of Evaluation, 1973.

Kapfer, M. B. (Ed.), Behavioral objectives: The position on the pendulum. Englewood Cliffs, N. J.: Educational Technology Publications, 1978.

Linn, R. L. Issues of reliability in measurement for competency-based programs. In M. A. Bunda & J. R. Sanders (Eds.) Practices and problems in competency-based measurement. Washington, D. C.: National Council on Measurement in Education, 1979. (a)

Linn, R. L. Issues of validity in measurement for competency-based programs. In M. A. Bunda & J. R. Sanders (Eds.) Practices and problems in competency-based measurement. Washington, D. C.: National Council on Measurement in Education, 1979. (b)

Linn, R. L. Issues of validity for criterion-referenced measures. Applied Psychological Measurement, 1980, 4, 547-561.

Linn, R. L. Criterion-referenced measurement: Psychometric problems. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, April 1981.

Livingston, S. A. Criterion-referenced applications of classical test theory. Journal of Educational Measurement, 1972, 9, 13-26.

Mager, R. F. Preparing instructional objectives. Palo Alto, CA: Reardon Publishers, 1962.

Mappus, L. L. Operationalizing broad scope objectives for measurement. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, April 1981.

Messick, S. Test validity and the ethics of assessment. American Psychologist, 1980, 35, 1012-1027.

Messick, S. Evidence and ethics in the evaluation of tests. Educational Researcher, 1981, 10 (9), 9-20.

Nitko, A. J. Distinguishing the many varieties of criterion-referenced tests. Review of Educational Research, 1980, 50, 461-485.

Popham, W. J. Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice-Hall, 1978.

Popham, W. J. & Lindheim, E. The practical side of criterion-referenced test development. NCME Measurement in Education, 1980, 10 (4).

Quelmallz, E. S. Criterion-referenced measurement: From the cognitive psychology view. A paper presented at the annual meeting of the American Educational Research Association, Los Angeles, April 1981.

Shoemaker, S. H. & Johnson, R. T. Assessing the construct validity of a criterion-referenced test: A nomological network approach. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, April 1981.

Subkoviak, M. J. Estimating reliability from a single administration of a criterion-referenced test. Journal of Educational Measurement, 1976, 13, 265-276.

Traub, R. E. & Rowley, G. L. Reliability of test scores and decisions. Applied Psychological Measurement, 1980, 4, 517-545.

Tennessee State Department of Education. The basic skills testing program in Tennessee. Nashville, TN: Author, 1981.

Wilcox, R. R. Determining the length of a criterion-referenced test. Applied Psychological Measurement, 1980, 4, 425-446.

Table 1

Number of Students Tested

GRADES	MATHEMATICS	SPELLING	LANGUAGE	READING
6	408	719	395	493
7	542	292	223	316
8	316	357	359	253
TOTAL	1266	1368	977	1062

Table 2

Average Difficulty of Subtests

GRADES	MATHEMATICS	SPELLING	LANGUAGE	READING
6	.58	.81	.89	.74
7	.64	.80	.87	.73
8	.72	.84	.89	.80

Table 3

Dependability Estimates (Brennan's Phi) for Subtests

GRADES	MATHEMATICS	SPELLING	LANGUAGE	READING
6	.92	.83	.89	.90
7	.93	.81	.87	.87
8	.93	.85	.89	.87

Table 4

Number of Objectives at Various Levels of Dependability

	MATHEMATICS	SPELLING	LANGUAGE	READING
Phi (.75)				
>.90	5	1	0	0
.85-.89	5	0	0	1
.70-.84	7	2	5	1
>.70	2	3	9	9
TOTAL	19	6	14	11

Table 5

Prophecy D Studies Summary

Phi (.75)≥.85

	Number of Objectives	Number of Items Required		Number of Objectives	Number of Items Required
MATHEMATICS	10	4	LANGUAGE	2	5
	2	5		1	6
	1	6		1	9
	1	7		1	10
	1	9		1	13
	1	10		1	15
	1	11		2	18
	1	14		5	>20
READING	1	4	SPELLING	1	4
	1	8		1	5
	1	13		1	7
	1	16		1	13
	2	17		1	15
	5	>20		1	>20

APPENDIX A

Tennessee Proficiency Test - Form IJ Objectives

MATHEMATICS OBJECTIVES

1. add four 3 digit, with regrouping
5. subtract fraction from mixed number, unlike denominators, with regrouping
7. divide a mixed number or a whole number by a fraction
8. add two or more decimal numbers, with regrouping
9. subtract two decimal numbers with regrouping
11. divide decimal number by whole number
13. place value, decimal
14. decimal fraction to percent
15. percent of a number
16. divide 4-digit number by a 1- or 2-digit number, with remainder
18. decimal equivalent
20. finding the perimeter
21. finding area
23. linear measurement - metric
24. writing a numeral for a word name
25. rounding off numbers
26. metric equivalency
28. compare four decimal numbers
29. determining average arithmetic mean

READING OBJECTIVES

63. synonym in context
64. antonym
65. words in context
68. story detail/time or setting
69. story detail/sequency of events
70. main idea
71. inference or conclusion
72. cause and effect
76. following directions
79. endings - suffixes
80. prefixes

SPELLING OBJECTIVES

31. base (root) word, with or without spelling change
32. spelling/phonetic substitution
33. spelling/missing letter
35. contractions
36. syllabication
38. doubling the final consonant

LANGUAGE OBJECTIVES

42. comma, phrases in a series
43. sentence punctuation
44. usage, irregular verbs
45. usage, nouns
47. parts of speech
48. subject-predicate agreement, inverted order
51. identifying simple subjects and verbs
52. identifying complete sentences
54. comma, in dates and addresses
55. quotations
58. agreement of possessive adjective
59. adjectives or irregular adjectives
60. verbs, principal parts
61. verbs, troublesome pairs

APPENDIX B

P values, Phi coefficients, and Phi $_{(.75)}$ coefficients

for each Objective by Grade for

Tennessee Proficiency Test - Form IJ

MATHEMATICS

Objective	GRADE 6 (N = 408) (P)	Phi	$Phi_{(.75)}$	GRADE 7 (N = 542) (P)	Phi	$Phi_{(.75)}$	GRADE 8 (N = 316) (P)	Phi	$Phi_{(.75)}$
1	(.91)	.53	.74	(.95)	.53	.83	(.95)	.37	.83
5	(.52)	.84	.88	(.63)	.89	.90	(.67)	.82	.83
7	(.45)	.84	.90	(.57)	.87	.89	(.61)	.86	.88
8	(.88)	.76	.80	(.92)	.66	.81	(.95)	.74	.90
9	(.80)	.70	.71	(.79)	.76	.79	(.91)	.74	.83
11	(.66)	.71	.72	(.78)	.69	.70	(.84)	.65	.69
13	(.80)	.85	.86	(.82)	.87	.87	(.83)	.85	.86
14	(.63)	.82	.84	(.70)	.82	.83	(.73)	.86	.86
15	(.40)	.74	.86	(.52)	.78	.84	(.60)	.83	.85
18	(.12)	.89	.98	(.25)	.94	.98	(.52)	.91	.93
20	(.33)	.94	.97	(.26)	.96	.98	(.61)	.96	.97
21	(.72)	.94	.94	(.76)	.94	.94	(.81)	.93	.93
23	(.74)	.75	.74	(.72)	.68	.67	(.72)	.70	.69
24	(.81)	.76	.77	(.85)	.75	.77	(.89)	.71	.78
25	(.53)	.55	.68	(.61)	.65	.70	(.65)	.58	.61
26	(.30)	.51	.85	(.30)	.61	.86	(.45)	.75	.84
28	(.33)	.92	.96	(.41)	.93	.96	(.53)	.93	.95
29	(.51)	.90	.92	(.60)	.93	.94	(.59)	.93	.94

SPELLING

	GRADE 6 (N = 719)			GRADE 7 (N = 292)			GRADE 8 (N = 357)		
Objective	(P)	Phi	Phi(.75)	(P)	Phi	Phi(.75)	(P)	Phi	Phi(.75)
31	(.97)	.70	.92	(.98)	.60	.95	(.98)	.69	.96
32	(.81)	.60	.61	(.81)	.63	.63	(.85)	.63	.67
33	(.83)	.77	.78	(.79)	.75	.75	(.85)	.81	.83
35	(.87)	.80	.83	(.87)	.78	.82	(.89)	.76	.82
36	(.71)	.60	.60	(.68)	.54	.56	(.74)	.64	.64
38	(.70)	.48	.50	(.69)	.47	.49	(.72)	.53	.53

	GRADE 6 (N = 395)			GRADE 7 (N = 223)			GRADE 8 (N = 359)		
Objective	(P)	Phi	$Phi_{(.75)}$	(P)	Phi	$Phi_{(.75)}$	(P)	Phi	$Phi_{(.75)}$
42	(.87)	.72	.77	(.87)	.81	.85	(.92)	.73	.84
43	(.91)	.77	.87	(.91)	.80	.87	(.91)	.79	.87
44	(.68)	.53	.60	(.62)	.59	.61	(.70)	.59	.59
45	(.67)	.37	.28	(.69)	.40	.35	(.70)	.50	.46
47	(.77)	.60	.59	(.73)	.57	.57	(.78)	.53	.51
48	(.73)	.14	-	(.75)	.26	-	(.74)	.20	-
51	(.64)	.28	-	(.67)	.28	-	(.73)	.30	.04
52	(.77)	.70	.70	(.79)	.72	.72	(.81)	.71	.71
54	(.60)	.65	.66	(.66)	.57	.63	(.73)	.62	.62
55	(.77)	.70	.70	(.73)	.70	.70	(.79)	.76	.77
58	(.57)	.08	-	(.56)	.14	.07	(.64)	.16	-
59	(.54)	.17	-	(.56)	.23	.24	(.59)	.32	-
60	(.91)	.60	.78	(.92)	.70	.81	(.94)	.58	.83
61	(.84)	.50	.59	(.86)	.46	.53	(.88)	.36	.54

READING

Objective	GRADE 6 (N = 493) (P)	Phi	$Phi_{(.75)}$	GRADE 7 (N = 316) (P)	Phi	$Phi_{(.75)}$	GRADE 8 (N = 253) (P)	Phi	$Phi_{(.75)}$
63	(.78)	.31	.08	(.80)	.41	.32	(.86)	.37	.43
64	(.59)	.67	.73	(.54)	.61	.71	(.62)	.77	.80
65	(.86)	.57	.64	(.85)	.50	.56	(.90)	.54	.69
68	(.75)	.49	.48	(.72)	.43	.42	(.81)	.46	.48
69	(.80)	.62	.62	(.77)	.55	.54	(.86)	.48	.58
70	(.61)	.37	.42	(.63)	.23	.23	(.68)	.18	.10
71	(.65)	.50	.52	(.63)	.43	.47	(.71)	.45	.40
72	(.74)	.60	.59	(.73)	.54	.53	(.79)	.56	.56
76	(.74)	.61	.59	(.72)	.61	.60	(.80)	.49	.49
79	(.91)	.80	.86	(.93)	.71	.85	(.94)	.61	.85
80	(.74)	.48	.44	(.70)	.34	.29	(.84)	.47	.53

APPENDIX C

P values, Range of P values, Range of Biserial Correlations, Phi coefficients, Phi $_{(.75)}$ coefficients and Number of Items Required for Phi $_{(.75)} \geq .85$ For Grades (6-8) for Tennessee Proficiency Test, Form IJ.

MATHEMATICS

GRADES 6-8

(N = 1266)

Objective	Range of P values	Range of Biserial Correlations	4 items Phi	4 items Phi(.75)	Number of items required Phi(.75) $\geq$ .85
1	.94 .92-.96	.54-.70	.51	.80	6
5	.60 .56-.63	.81-.87	.86	.88	4
7	.54 .51-.57	.81-.87	.86	.89	4
8	.92 .88-.	.69-.79	.72	.83	5
9	.85 .82-.89	.74-.77	.74	.77	7
11	.76 .66-.81	.73-.76	.71	.71	10
13	.82 .79-.85	.82-.86	.86	.86	4

C2

Objective	Range of P values	Range of Biserial Correlations	4 items Phi	4 items Phi(.75)	Number of items required Phi(.75) $\geq$ .85
14	.68 .61-.72	.81-.84	.84	.84	5
15	.50 .42-.55	.74-.82	.79	.85	4
16	.70 .60-.79	.62-.73	.62	.62	14
18	.28 .25-.31	.90-.93	.93	.97	4
20	.37 .36-.38	.92-.96	.96	.98	4
21	.76 .73-.79	.90-.96	.94	.94	4
23	.72 .59-.87	.67-.82	.71	.70	9
24	.84 .81-.86	.69-.80	.75	.77	7
25	.59 .58-.65	.66-.76	.61	68	11

C3

Objective	Range of P values	Range of Biserial Correlations	4 items Phi	4 items $Phi_{(.75)}$	Number of items required $Phi_{(.75)} \geq .85$
26	.35 .32-.40	.55-.77	.64	.86	4
28	.41 .40-.43	.90-.92	.93	.96	4
29	.57 .57-.58	.89-.91	.92	.93	4

SPELLING

GRADES 6-8

(N - 1368)

Objective	Range of P values	Range of Biserial Correlations	4 items Phi	4 items $Phi_{(.75)}$	Number of items required $Phi_{(.75)} \geq .85$
31	.97 .95-.98	.66-.80	.69	.94	4
32	.82 .75-.92	.62-.72	.61	.63	13
33	.82 .77-.89	.76-.80	.78	.79	7
35	.87 .85-.90	.75-.81	.78	.82	5
36	.71 .64-.73	.62-.70	.60	.60	15
38	.70 .65-.75	.56-.69	.49	.50	>20

LANGUAGE

GRADES 6-8

(N = 977)

Objective	Range of P values	Range of Biserial Correlations	4 items Phi	4 items Phi(.75)	Number of items required Phi(.75) ≥ .85
42	.89 .86-.91	.72-.80	.77	.83	5
43	.91 .88-.93	.73-.80	.76	.84	5
44	.67 .57-.77	.62-.74	.58	.69	15
45	.68 .42-.90	.49-.70	.43	.38	>20
47	.76 .64-.85	.63-.73	.57	.56	18
48	.74 .26-.93	.57-.62	.21	-	>20
51	.68 .24-.86	.43-.72	.30	.02	>20

Objective	Range of P values	Range of Biserial Correlations	4 items Phi	4 items Phi(.75)	Number of items required Phi(.75) ≥ .85
52	.79 .69-.87	.72-.77	.71	.71	10
54	.66 .53-.78	.63-.72	.62	.64	13
55	.77 .74-.81	.69-.80	.73	.73	9
58	.60 .22-.93	.40-.72	.15	.01	>20
59	.56 .19-.90	.43-.75	.25	.22	>20
60	.92 .91-.94	.68-.72	.64	.81	6
61	.86 .80-.92	.54-.66	.44	.55	18

READING

GRADES 6-8

(N = 1062)

Objective	Range of P values	Range of Biserial Correlations	4 items Phi	4 items Phi(.75)	Number of items required Phi(.75) ≥ .85
63	.81 .49-.95	.49-.75	.36	.25	>20
64	.59 .47-.70	.63-.79	.69	.74	8
65	.87 .75-.94	.59-.74	.54	.64	13
68	.76 .67-.83	.59-.70	.47	.47	>20
69	.81 .74-.89	.60-.71	.58	.59	16
70	.63 .44-.72	.48-.64	.29	.31	>20
71	.67 .42-.78	.57-.70	.48	.48	>.20

Objective	Range of P values	Range of Biserial Correlations	4 items Phi	4 items $Phi_{(.75)}$	Number of items required $Phi_{(.75)} \geq .85$
72	.75 .69-.86	.60-.71	.58	.57	17
76	.75 .67-.84	.61-.71	.58	.57	17
79	.92 .91-.94	.73-.77	.75	.86	4
80	.76 .60-.90	.59-.66	.46	.43	>20